国外搜索引擎检索效能研究述评*

马费成　望俊成　吴克文　邱　璇

摘　要　在网络搜索引擎的使用中,搜索引擎的检索效能成为影响用户信息获取效果和搜索引擎服务质量的重要因素。目前,国外的相关研究主要采取实验的方法,从用户体验角度出发评价搜索引擎的检索效能,主要步骤包括确定信息需求、选择搜索引擎、评价结果文档相关度以及确定测度指标。最常用的测度指标是查全率和查准率。此外,影响用户检索效能的指标还有搜索引擎返回结果文档的排序质量、重复度,而索引的数量、用户满意度等指标都会影响用户使用的效果。无论是从搜索引擎的用户使用角度,还是用户评价角度,"用户参与"的模式是最贴近检索现实的。表1。参考文献27。

关键词　信息检索　搜索引擎　检索效能　对比试验　述评
分类号　G354

ABSTRACT　Retrieval effectiveness of search engines becomes an important factor which affects users' effects of acquiring information and search engines' quality of service. At present, some foreign researches adopt experimental method to evaluate retrieval effectiveness of search engines from user's experience. Main steps include ascertaining information needs, choosing search engines, evaluating documents' correlation measures, and ascertaining measuring index. Recall and the accuracy ratio are used most. Beside, there are some other indexes that affect users' retrieval effectiveness, such as sorting quality of returned documents and multiplicity. But, number of index and customer satisfaction also can influence users' effects. The mode of user's participation is the closest to reality retrieval either from the angle of search engine users, or users' evaluation. 1 tab. 27 refs.

KEY WORDS　Information retrieval. Search engine. Retrieval effectiveness. Comparative trial. Systematic review.
CLASS NUMBER　G354

1　引言

网络信息井喷式增长使得信息获取与用户需求之间的矛盾愈发突出,如何快速、准确地从网络上获取用户所需要的信息成为信息服务提供者不断努力的方向。搜索引擎,作为一种在网络上对信息进行收集、提取、组织、处理并提供检索服务的工具,极大地改善了用户的网络信息获取过程和效果。信息检索已经成为三大基础应用之一。

在搜索引擎的使用过程中,返回的是根据查询语句得到的拥有数以万计结果的文档集。一般情况下,用户只会浏览前几条数据,不可能系统地浏览所有查询结果[1]。因此,搜索引擎排序的质量成为影响用户信息获取效果和搜索引擎服务质量的重要因素。对搜索引擎结果进行准确而有意义的评价不仅可以引导用户作出合理的选择,而且可以帮助搜索引擎服务商提高服务标准[2-3]。本文利用系统综述的方法,对国外搜索引擎检索效能评价的相关研究进行全面评述,旨在梳理主流的搜索引擎评价方法、评价流程及评价指标。

本文的资料来源主要是公开发表的英文文

* 本文为国家自然科学基金项目"基于Web2.0的信息自组织与有序化研究"(项目编号:70773086)系列研究成果之一。

献。首先通过对 retrieval effectiveness & search 以及 search engine & measure 两个查询语句在两个专业引文数据库——科学引文索引扩展板(SCIE)和社会科学引文索引(SSCI)中进行检索。通过筛选,得到 Gordon 和 Pathak(1999),P. Jason(2008)等 4 篇最相关的文献,再对其参考文献进行甄别和检索,最终获得与搜索引擎效能评价研究相关的文献 30 篇。本述评就是在对这些文献仔细阅读、分析和抽取的基础上写成的。

2　搜索引擎评价方法

搜索引擎主要有四种类型:目录式搜索引擎、全文检索式搜索引擎、元搜索引擎[4]和基于大众分类的搜索引擎[5]。其中针对网络搜索引擎展开的检索效能评价研究的方式主要有两种:一种是从搜索引擎本身特征入手,侧重于一些直观的观测数据,如搜索引擎的响应速度、稳定性、易用性、界面设计等特征。尽管这种直接观测得到的相关指标,能够在用户选择搜索引擎的过程中提供一些有用的信息,但是这些指标并不能直接让用户认识到哪个搜索引擎更加有效。第二种是采取实验的方法对比研究,从用户体验角度出发评价搜索引擎的检索效能,主要步骤包括了确定信息需求、选择搜索引擎、评价结果文档相关度以及确定测度指标。

采用对比实验方法对搜索引擎检索效率的研究较多。Gordon 和 Pathak 在 1999 年就比较了 12 个早前的类似研究,但使用恰当的实验设计和评估方法的研究并不多见。Marchionini 使用 5 个查询语句完成了对 3 个搜索引擎的检索效能比较研究,没有发现它们在数学意义上的显著区别[6];Leighton 和 Srivastava 使用 15 个查询语句对 5 个搜索引擎进行了比较研究,发现其中 3 个搜索引擎的检索效能明显高于另外两个[7]。该研究首次使用了以每个搜索引擎返回的前 20 条检索结果为研究对象的方案。早期类似研究中,Gordon 和 Pathak 的研究是最全面的[8]。后续的许多研究大多沿袭了 Gordon 和 Pathak 的研究框架[9-11]。4 个类似实验的详细对比见表 1。

表 1　同类研究的相关指标对比

指标	Gordon 和 Pathak (1999 年)	Hawking 等 (2001 年)	Can Nuray 等 (2003 年)	P. Jason (2008 年)
信息需求提供者	学院教师	日志文件	学生和教授	硕士研究生
查询语句提交者	熟练的检索者	同上	同上	同上
相关性判断者	同一批学院教师	助理研究员	同上	同上
参与人数	33	6	19	34
人均检索次数	1	9	1-2	3-4
总检索次数	33	54	25	103
研究的搜索引擎数	8	20	8	8
每个搜索引擎获取的条目数	20	20	20	20
人均判断的条目数	160	3600	160 或 320	大约 160
相关性判断方法	四值判断	二值判断	二值判断	二值判断
查准率测算方法	P(1-5),P(1-10) P(5-10),P(15-20)	P(2),P(1-5) P(5),P(20)	P(10),P(20)	P(20),P(1-5)

续表

指标	Gordon 和 Pathak（1999 年）	Hawking 等（2001 年）	Can Nuray 等（2003 年）	P. Jason（2008 年）
查全率测算方法	R(15 -20),R(15 -25) R(40 -60),R(90 -110), R(180 -200)	无	R(10),R(20)	R(20),R(1 -5)
结果重叠率测度	有	无	无	有
结果排序质量测度	无	无	无	无

注:P 代表相对查准率,R 代表相对查全率。

3 确定信息需求

Mizzaro[12]将检索得到的结果文档划归到相应的信息需求类别。Gordon 和 Pathak 在 1999 年的研究中则进一步针对如何更好地进行对比实验研究,提出了七点建议。其中有三点是针对信息需求提出的:首先,信息需求必须是真实的;其次,为了补充查询语句,信息需求应该尽可能详细地被描述出来;再者,信息需求应该包含不同的主题,以及不同类型的预期结果。Hawking 等[13]在研究中专门针对信息需求进行了分类,按照预期的检索结果类型提出信息需求分为以下四类:①事实类信息需求,需要的结果是一段简短的事实型描述;②精准类信息需求,需要的结果是某个特定的文档、网页或者网站;③兴趣类信息需求,需要的结果是满足用户某类兴趣的一批相关文档,这些兴趣可以是学术类、新闻类或者其他类别的;④需要的结果是与选择标准相匹配的所有文档,这个匹配是语法层面,也可以是语义层面。由于第四类信息需求很难评价结果是否与需求相关,因此后续的研究大多只是以前三类信息需求为对象。P. Jason Morrison[11]在 Hawking 前三个信息需求的基础上,另外扩展了三类:新闻类,普通类和娱乐类。

4 选择搜索引擎

互联网上有着数以万计各式各样的搜索引擎,但是占据行业统治地位的却屈指可数。其中被用于研究评价的搜索引擎主要是 Google、AltaVista 和 Yahoo 三种。选择 Google 是因为它是当前最大的公用搜索引擎,它的 PageRank 算法是基于链接的排序算法中的佼佼者。AltaVista 是最老的 Web 搜索引擎之一,且在数据库容量方面具有领先优势,是网络计量学研究最常用的搜索引擎。Yahoo 是老牌的目录式搜索服务提供商,精准的目录式信息分类成为其服务特色,在多种类型的搜索引擎比较实验中常被选为目录式搜索引擎的代表。

对于选择搜索引擎,Gordon 和 Pathak 在 1999 年的研究中提出了两点建议:首先,大多数的主流搜索引擎,都应该被纳入研究范畴;其次,某个搜索引擎的"个别特性"应该被考虑,即使被提交的实际查询不能被其它搜索引擎识别。在搜索引擎的数量方面,从实验成本出发,应选择 10 个以下,如 Leighton 和 Srivastava[7]选择了 5 个搜索引擎,Gordon 和 Pathak[8]、Can Nuray 等[10]和 P. Jason Morrison[11]都选择了 8 个搜索引擎,而 Hawking 等[9]的选择达到 20 个。搜索引擎的类别方面,大多数研究都是针对全文式搜索引擎和目录式搜索引擎,也有的研究是在综合搜索引擎和垂直搜索引擎之间,以及综合搜索引擎和元搜索引擎之间展开,而 P. Jason Morrison[11]则首次将大众分类网站引入到搜索引擎检索效率对比研究的范畴。

目前,国内的搜索引擎市场已经逐步形成一批市场占有率较高的主流搜索引擎,全文检索搜索引擎如百度、Google;目录式搜索引擎如

搜狗目录,Google 目录;垂直搜索引擎如百度新闻、新浪新闻、Google 学术、腾讯娱乐搜搜等。这给国内学者提供了进行此类研究的基础,但是需要注意的是,在确定搜索引擎时,数量的把握需要兼顾实验目的和成本,不同类别搜索引擎之间的比较也要考虑到合理的信息需求类别。

5　评价文档相关性

评价结果文档相关性主要有用户评价与程序自动评价两种方法。用户评价文档相关性,基本上采取"打分判断"的方法。不同的研究采取不同的"分值类型",大多数研究采取的是"二值判断"(相关,不相关),Voorhees 和 Harman[14]、Hawking 等[9]以及 P. Jason Morrison[11]都是采用二值来判断结果文档相关与否。也有一些研究采取了"多值判断"来进一步区分结果文档相关度的强弱。Chu 和 Rosenthal[15]在研究中采用了"三值判断"(相关,一般相关,不相关),Gordon 和 Pathak[8]采取的是"四值判断"(强相关,弱相关,较不相关,很不相关),而 Ding 和 Marchionini[6]则将相关度的判断进一步细化,采用了"六值判断"(完全相关,强相关,弱相关,较不相关,很不相关,完全不相关)。

并非所有的搜索引擎文献都是使用人的判断作为评价基础[4]。Courtois 和 Berry[16]研究了 5 个搜索引擎返回的 100 个检索结果,利用计算机程序自动检测地址和相似度等指标判断相关性。Can Nuray 等[10]根据相关性人工评价方法,提出 AWSEEM 网络搜索引擎相关性自动评价方法。该研究发现自动评价方法与人工评价方法之间存在着统计上的强相关联系。Kawaguchi[17]也采用了类似的自动评价方法;Courtois 和 Berry[16]则利用"自编程序"来判断结果的相关性。

如果采用人工判断相关性的方法,判断者本身就成为了一个实验变量,那么由谁来判断各个搜索引擎检索出来的结果文档的相关性?是领域专家、信息需求提出者自己,还是研究者? Voorhees 和 Harman[14]指出应当由专家来判断;Chu 和 Rosenthal[15]提出由研究者自己来判断;而 Gorden 和 Pathak[8]则强调相关性判断必须由提出原始信息需求的个人完成,这样能更加接近检索行为的现实情况,体现用户体验的真实反馈。

6　相关测度指标

6.1　查全率和查准率

从经典的 Cranfield 项目[18]到现在的文本检索会议(Text Retrieval Conference,TREC),信息检索系统评价已经建立起一套衡量指标,这套指标中最常用的就是查全率和查准率。查全率是指检出的相关文档量与检索系统中相关文档总量的比率;查准率指检出的相关文档量与检出文档总量的比率。

该套指标适合类似于 Cranfield 项目之类的小型实验室实验,但随着文档集合的增长,变得越来越不适合,在 Web 环境下尤其突出。以查全率为例,很难统计出一个检索词下互联网中所有相关文档的总量。因此,许多学者认为查全率并不适合作为一个评价指标。Chu 和 Rosenthal 的 Web 检索实验甚至屏蔽了查全率这个指标。Gwizdka 和 Chignell 在 1999 年的研究中也因为同样原因没有将查全率加入到搜索引擎评价体系中来;而是在此基础上衍生出其他多种评价指标,其中相对查全率最为典型。以检索出来的相关文档总数为基数展开研究,在不同阈值下求出相对查全率。如:同时对 n 个搜索引擎 Si(S1,S2,…,Sn)提交同一查询,各搜索引擎返回的文档数量分别为 Ai (A1,A2,…,An),而其中与用户需求相关的文档数量为 Bi (B1,B2,…,Bn)。那么这 n 个搜索引擎检索出来的相关文档总数量 $B = \sum_{i=1}^{N} Bi - C$(C 表示使用不同搜索引擎检索的结果中重复的文档数量),某个特定搜索引擎 Si 的相对查全率 $Ri = Bi/(\sum_{i=1}^{N} Bi - C)$。不同阈值 t 下的相对查全率,则是以每个搜索引擎返回的前 t 条结果为对象来展开研究。类似的,相对查准率则是由某个特定搜索引擎检索出的相关文档数量与该搜索引擎检索出的所有文档数量的比值决定,特定搜索引擎

Si 的相对查准率 Pi = Bi/Ai。

查准率和查全率已成为搜索引擎检索效能评价研究中的两个经典指标,但国外学者的研究从来没有停止对这两个指标的不断验证和完善,尤其是查全率这个指标,从查全率到相对查全率,再到不同阈值下的相对查全率,力求最大可能地挖掘这个指标对检索效能的评价作用。在综述的过程中,笔者认为在实践中这两个指标还可以进一步完善,如针对 TopN 的结果文档集测度各个搜索引擎返回的准确率和完备率,因为最相关文档才是用户真正关心的资源。

6.2 结果文档的排序质量

除了高的查准率和查全率,结果文档排序的质量好坏对于提高网络搜索引擎检索质量和用户体验效果也很重要。如果没有好的排序,再精确、再完备的结果也会因为排列混乱而得不到用户的青睐。Courtois 和 Berry[16] 专门针对网络搜索引擎结果排序进行研究,指出结果排序在用户对网络搜索引擎的满意度以及检索相关文档的成功率方面有重要影响。搜索引擎结果文档排序的质量也是评价检索效能的一个很好的指标。

然而,对于网络搜索引擎结果文档排序质量的研究并不是很多。Su, Chen 和 Dong[19] 使用"五值判断"对同一查询下各搜索引擎返回的结果文档进行人工排序,然后分别以每个搜索引擎返回的结果文档为对象,计算这些结果文档的系统排序与人工排序这两个序列之间的 spearman 相关系数。该系数的取值范围为[-1,1],1 表示排序完全一致,-1 表示顺序完全相反。相关系数越大,表明该搜索引擎的系统排序质量越接近用户对文档相关度的评价。Guizdka 和 Chignell 在 1999 年的研究中认为检索结果的排序是结果展示的重要方面,为了评价排序的质量,他们在"四值判断"人工排序的基础上,引入了"偏微分查准率"这一新指标来测度搜索引擎的排序质量。

在检索效能评价过程中,计算的变量取决于相关性判断的实验方式。许多研究都使用了二值相关性判断[9,17,11,20]。由于注意到相关性存在"程度"的差别,因此出现了多值相关判断。Chu 和 Rosenthal[15] 使用三值相关判断,Gordon 和 Patha[8] 使用四值相关判断。Su, Chen 和 Dong[19] 在对搜索引擎检索结果排序时使用五值判断。Ding 和 Marchionini[6] 在计算三种不同类型的精确度时使用了六值判断。用户对结果文档"N 值判断"的操作,实际上是对文档相关性采取的一种不连续的评分方法,很容易出现多个文档获取相同的分数,在排序上难分上下,所以这种"N 值判断"的评分系统在测度结果排序的质量上并不是十分有效。针对此问题,Liwen Vaughan[3] 提出了一种连续的排序方法,要求实验者对结果文档进行全面的比较后,根据各自对结果文档与信息需求相关程度强弱的判断,对该集合中所有文档排列序次,然后再计算系统排序与人工排序之间的相关系数作为衡量结果排序质量的一个指标。

结果文档排序质量的相关系数,无疑是一个很好的测度指标,不过该指标的测度也存在一定的难度,如何获取结果文档在人工排序中的序次是一个值得深入探讨的问题。如果采用离散式的"N 值判断"评分系统会导致许多文档的序次一样,这就没有达到区分检索效能高低的作用;而如果采用连续的排序方式,对不相关结果文档的人工序次又该如何赋值?只有人工序次的问题得到很好的解决,结果文档排序质量系数才能成为一个有效的测度指标。

6.3 搜索引擎之间结果文档重复度

Lawrence 和 Giles[21] 在 Science 上发表的报告指出单一的搜索引擎只能覆盖整个网络的 3% - 34% 的内容。因此,对同一个查询,各搜索引擎返回的结果集必然存在着重叠。有的结果文档仅被一个搜索引擎返回,而有的同时被多个搜索引擎返回,也就意味着结果文档具有不同的重复度。统计不同重复度下单一文档数量和相关文档数量的分布情况,能够从一定程度上反映该搜索引擎检索主流文档的能力。同时,针对搜索引擎两两之间文档重叠情况,计算得到的"结果文档重复率"和"相关文档重复率"可以作为查准率和查全率的补充。如搜索引擎

S1 和 S2 针对同一查询返回的结果文档数量分别为 A1 和 A2,两个结果集重叠的数量为 C1;其中返回的相关文档数量分别为 B1 和 B2,重叠的数量为 C2,那么这两个搜索引擎的结果文档重复率为 C1/(A1 + A2),相关文档重复率为 C2/(B1 + B2)。

Ding 和 Marchionini[6] 首次对相同查询下,不同搜索引擎的检索结果之间的重叠问题进行了探讨;Bharat 和 Broder[22] 对 HotBot、Alta Vista、Excite 和 InfoSeek 进行了研究,估算出相同查询下这 4 个搜索引擎的检索结果的重叠率仅为 1.4%。Gorden 和 Pathak[8] 在不同的文档阈值下(DCV,document cut-off value),测度了 5 个搜索引擎之间的结果文档重叠率,发现相关结果中的 93% 仅仅在一个搜索引擎中出现,而且这个比例即使是在较高的 DCV 下也表现得十分稳定。这表明大量相关结果文档广泛分布于各个搜索引擎。

2000 年以后,许多研究针对不同搜索引擎进行了大量的实证分析[1,10,17,22-25],进一步证实了相同检索词下搜索引擎返回结果之间低重叠现象的存在。其中 Spink 和 Jansen 等[26] 专门针对相同检索词下搜索引擎之间的检索结果重叠问题进行了研究,作者在 4 个搜索引擎以及一个元数据搜索引擎之间进行了一个大范围的研究,目的在于考证小型实验中的重叠现象在大型实验中又有着怎样的表现,最后得出的结论是低重叠率在该大型实验中表现得依然明显,进而指出元搜索引擎可以很好地解决单一搜索引擎存在检索偏好不足的缺陷。P. Jason Morrison[11] 在目录式搜索引擎、全文式搜索引擎和大众分类网站这三类搜索引擎执行效率的研究实验中,对相同检索词下这三类搜索引擎的结果重叠情况进行了研究,最后指出目录式搜索引擎和全文式搜索引擎之间的重叠情况要明显高于这二者分别与大众类搜索引擎之间的重叠率。

国外的此类对比实验大多采用了结果重叠率这一指标,笔者认为这只能从相对的角度反映两个搜索引擎之间的检索结果相似情况,难以同查准率、查全率一样成为独立的效能评价指标。然而,如果以某一个搜索引擎为标杆,来计算和比较其他各搜索引擎同该标杆搜索引擎的结果重叠情况,还是能够从一定的角度来反映搜索引擎之间检索效能高低的。但是很难找到一个好的标杆,因为即使检索效能表现相对优秀的 Google、百度,这二者之间的重叠率也很低。因此结果重复率这一指标还有待进一步改进。

6.4　其它的测度指标

在网络搜索引擎检索效能的研究中,除了查准率、查全率、结果排序质量和结果重叠度指标外,还存在许多其他的测度指标和方法,如“索引的数量”、“用户满意度”等。

索引文档的数量是指如 Google 公布互联网最新索引数量达到以兆计(10^12 个)的网页数量。Hawking 等[9] 通过研究发现索引数量与检索效能之间并不存在明显的正相关。而 Hawking 和 Robertson 于 2003 年再次进行类似研究,又指出索引数量可以提高检索效能。Beg[27] 定义了一个“用户满意度”的测度指标,该研究中参与者并不直接判断检索结果的相关性,而是通过监测用户大量的隐性因素,如用户判断条目的次序,用户判断该文件花费的时间,用户是否打印、复制或者保存该文档,以及其他行为。然后研究者对这些隐性因素分别赋以权重,建立数学模型,形成综合性的指标来评价用户对文档的满意度。

目前以查准率为代表的常规指标基本成熟,但是对于其它的测度指标还有很大的开发空间,应该开发和定义出更能反映用户真实检索感受的指标。例如“用户满意度”这个指标,就是从搜索用户本身的行为入手来评价搜索引擎检索效能的优劣。这种利用用户日常检索行为的反馈开发的评价指标更能接近现实,其评价的指导意义也更加明显。

7　结语

对于搜索引擎检索效能的评价,学术界已经开展了多年的研究,本文回顾了国外学者利

用对比实验进行搜索引擎检索效能研究的实验过程,梳理了对比实验需要注意的细节,包括:①实验参与者尽量保持一致,即信息需求提供者、查询语句提交者和相关性判断者尽可能地保持一致,这样才能保证相关性的判断是从需求用户的角度出发。虽然提交的查询语句不一定专业,但它更加接近真实的网络搜索行为;②采用多种相关性判断方法,为了更好地区分结果文档相关性的强弱,多值判断比二值判断更加有效。但是离散式的判断方法依然存在难以较好地区分相关性的问题,因此需要引入诸如全排序的连续型判断方法和程序自动判断的方法;③合适选择和解释测度指标。实际上可供选择的测度指标已经很多,但是并非每个对比实验都需要将所有的测度指标全部计算,需要结合实验的目的和可供比较的实验结果来选择,同时进行合理的解释。

国外利用对比实验进行搜索引擎检索效能的评价研究,尽管已经形成一定的研究范式,仍然有许多需要完善和改进的环节。结合笔者的实践,总结归纳出以下结论,希望能对后续同类研究提供方法上的指导。首先,确定信息需求需要进一步细化,这样可以充分地考察某一个搜索引擎在满足哪一类信息需求方面更有优势;其次,关于元搜索引擎及新兴的基于大众分类的搜索引擎的研究还不够深入;再者,此类评价研究已经形成了一批成熟的评价指标,开发出了能够更好反映用户真实检索感受、更明显地区分检索效能高低的测度指标,构建了合理的评价指标体系,这些为后续的研究提供了借鉴;最后,无论是从搜索引擎的用户使用角度,或者是用户评价角度,"用户参与"的模式无疑是最贴近检索现实的,但是在引入用户参与的过程中,也要考虑到参与成本问题。

参考文献:

[1] Bar-Ilan, J. Comparing rankings of search results on the Web[J]. Information Processing and Management, 2004(41): 1511 - 1519.

[2] David Hawking, Nick Craswell. Measuring search engine quality[J]. Information Retrieval, 2001(4): 33 - 59.

[3] Liwen Vaughan. New measurements for search engine evaluation proposed and tested[J]. Information Processing and Management, 2004(40): 677 - 691.

[4] Bernard J, Jansen. A Review of web searching studies and a Framework for future research[J]. Journal of the American Society for Information Science and Technology, 2001(52): 235 - 246.

[5] Andreas Hotho, Robert Jaschke, Christoph Schmitz, Gerd Stumme (2006). Information retrieval in folksonomies: search and ranking[J]. Lecture Notes in Computer Science, 2006(4011): 411 - 426.

[6] Ding, W, Marchionini, G. A comparative study of Web search service performance[R]. In Proceedings of the 59th annual meeting of the American Society for Information Science, 1996:136 - 142.

[7] Leighton, H. V, Srivastava, J. First 20 precision among World Wide Web search services (search engines)[J]. Journal of the American Society for Information Science, 1999, 50(10): 882 - 889.

[8] Micheal Gordon, Praveen Pathak. Finding information on the World Wide Web:the retrieval effectiveness of search engines[J]. Information Processing and Management, 1999(35): 141 - 180.

[9] Hawking, D, Craswell, N, Bailey, P, Griffihs, K. Measuring search engine quality[J]. Information Retrieval, 2001(4): 33 - 59.

[10] Can F, Nuray R, Sevdik A. B. Automatic performance evaluation of Web search engines[J]. Information Processing and Management, 2004(40): 495 - 514.

[11] P. Jason Morrison. Tagging and searching:Search retrieval effectiveness of folksonomies on the World Wide Web. Information Processing and Management, 2008(44): 1562 - 1579.

[12] Mizzaro S. Relevance: The whole history[J]. Journal of the American Society for Information Science, 1997(48): 810 - 832.

[13] Hawking D, Robertson, S. On collection size and retrieval effectiveness[J]. Information Retrieval, 2003, 6(1): 99 - 105.

[14] Voorhees, E. Variations in relevance judgements

and the measurement of retrieval effectiveness. Information Processing and Management, 2000, 36(5): 697 - 716.

[15] Chu H, Rosenthal, M. Search engines for the World Wide Web: a comparative study and evaluation methodology[R]. In Proceedings of the 59th annual meeting of the American Society for Information Science, 1996: 127 - 135.

[16] Courtois M. P, Berry M. W. Results ranking in Web search engines[J]. Online, 1999, 23(3): 39 - 46.

[17] Mowshowitz, A., Kawaguchi, A. Assessing bias in search engines[J]. Information Processing and Management, 2002, 35(2): 141 - 156.

[18] Cleverdon, C. W. On the inverse relationship of recall and precision[J]. Journal of Documentation, 1972, 28(3): 195 - 201.

[19] Su L. T, Chen H, Dong X. Evaluation of Web-based search engines from the end-user's perspective: a pilot study[R]. In Proceedings of the 61st annual meeting of the American Society for Information Science, 1998: 348 - 361.

[20] D. Hawking, N. Craswell, P. Thistlewaite, D. Harman. Results and challenges in Web search evaluation[R], In Proceedings 8th International World Wide Web Conference (WWW8), 1999(5): 1321 - 1330.

[21] S. Lawrence, C. L. Giles, Searching the World Wide Web[J], Science, 1998, 5360 (280): 98 - 100.

[22] Bharat K, Broder A. A technique for measuring the relative size and overlap of public Web search engines[J]. Computer Networks and ISDN Systems, 1998, 30(1 - 7): 379 - 388.

[23] Nicholson, S. Raising reliability of Web search tool research through replication and chaos theory[J]. Journal of the American Society for Information Science, 2000,51(8): 724 - 729.

[24] Hood, W, W. Wilson, C. S. Overlap in bibliographic databases[J]. Journal of the American Society for Information Science and Technology, 2001, 54(12): 1091 - 1103.

[25] Ferreira, J, da Silva, A. R, Delgado, J. Does overlap mean relevance[R]. In Proceedings of WWW/Internet 2004 (IADIS) conference, 2004(10).

[26] Spink A, Jansen B, Blakely C. Koshman, S. A study of results overlap and uniqueness among major Web search engines[J]. Information Processing and Management, 2006, 42(5): 1379 - 1391.

[27] Beg M. A subjective measure of Web search quality[J]. Information Sciences, 2005, 169(3 - 4): 365 - 381.

马费成　武汉大学信息资源研究中心主任,教授,博士生导师。通讯地址:武汉。邮编430072。

望俊成　武汉大学信息管理学院博士研究生。通讯地址同上。

吴克文　邱　璇　武汉大学信息管理学院硕士研究生。通讯地址同上。

(收稿日期:2008-12-31;修回日期:2009-02-22)